DOCUMENT RESUME

ED 262 084 TM 850 565

AUTHOR Phillips, Gary W.
TITLE Statistical Equating of Direct Writing Assessment.
PUB DATE 85
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Equated Scores; Equations (Mathematics); *Error of Measurement; *Essay Tests; Field Tests; Holistic Evaluation; *Statistical Analysis; *Writing Evaluation
IDENTIFIERS Linear Equating Method; *Maryland Functional Writing Test; *Rasch Model

ABSTRACT

This paper provides empirical data on two approaches to statistically equate scores derived from the direct assessment of writing. These methods are linear equating and equating based on the general polychotomous form of the Rasch model. Data from the Maryland Functional Writing Test are used to equate scores obtained from two prompts given in 1984 to the scores obtained from two different prompts administered in 1983. Two 1983 writing prompts were administered to 351 examinees, while two 1984 prompts were administered to 347 examinees. The results suggest two conclusions: (1) the effects of statistical equating depends in part on where within the score range the passing score happens to be; and (2) different results are obtained with different equating methodologies. It is recommended that the equating of direct writing assessments be accomplished with the general polychotomous form of the Rasch model. All the advantages of the Rasch model for multiple choice tests, potentially, may be realized in direct writing assessment. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262084

STATISTICAL EQUATING

OF

DIRECT WRITING ASSESSMENT

Gary W. Phillips

Maryland State Department of Education

1985

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Phillips, Gary W.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 505

Paper presented at the annual convention of the American Educational Research Association, Chicago, 1985.

INTRODUCTION*

With the advent of large scale proficiency testing many school districts and state agencies are making certificatioin decisions, in part, based on direct writing tests. Alternate forms of the same test must be equated in order to maintain score scale comparability. This is especially important for testing programs which monitor longitudinal trends and/or must maintain an invariant passing standard.

Pre-equating methods are often used to improve score scale comparability across alternate forms. Pre-equating often takes the form of: (1) choosing prompts from a field test with similar score distributions (this is analogous to p-value equating on multiple choice tests); and (2) maintaining constant training standards, rubrics and test specifications (Williams, 1983). In spite of these pre-equating strategies, the lessons from multiple choice testing have shown that some type of statistical equating (i.e., adjustment to the score scale) is often required.

*Some portions of this paper were presented at the Education Commission of States annual conference on large-scale assessments, Boulder, Colorado, 1984.

This paper provides empirical data on two approaches to statistically equate scores derived from the direct assessment of writing. These methods are: (1) linear equating; and (2) equating based on the general polychotomous form of the Rasch model. Equipercentile equating was not performed because it was felt that not enough score point intervals were present in the score range. In both cases, data were obtained using the random groups design (Angoff, 1971). In this design test forms Y and X (with no anchor items in common) are given to two randomly equivalent groups A and B.

Linear Equating

In linear equating the scores from two test forms, Y and X, are said to be equated if they correspond to equal standard-score deviates. This method of equating is appropriate when test form distributions differ only with respect to their means and standard deviations. Under these conditions test form X given to group B may be equated to test form Y given to group A (given they are equally reliable after equating) via the equation

$$Y^{*}_{A} = \frac{S_{YA}}{S_{XB}} X_{B} + \bar{Y}_{A} - \frac{S_{YA}}{S_{XB}} \bar{X}_{B} \quad . \qquad (1)$$

This conversion equation forms a straight line with intercept and slope equal to $\bar{Y}_{A} - (S_{YA}/S_{XB})\bar{X}_{B}$ and S_{YA}/S_{XB}, respectively (Angoff, 1971).

The standard error of linear equating is the standard deviation of Y^* scores for a fixed value of X. According to Angoff (1971) the standard error of equating for the random groups design is

$$SE(Y^*) = \left[\frac{2S_y^2(Z_x^2+2)}{N_A+N_B}\right]^{1/2} \qquad (2)$$

where $Zx=(X-\bar{X})/Sx$. The $SE(Y^*)$ is smallest at $\bar{X}$ and increases in magnitude as X deviates from $\bar{X}$.

Polychotomous Rasch Model Equating

In the general polychotomous form of the Rasch model (also called the partial credit model) the possible levels of performance on item i are labeled $o,1,\ldots,m_i$. The ith item is conceptualized as a sequence of m_i steps. The model for the Xth step on item i is

$$P_{nix} = \frac{\exp(\beta_n-\delta_{ix})}{1+\exp(\beta_n-\delta_{ix})} \qquad \text{where } X = 1,2,..,m_i \qquad (3)$$

In equation 3 P_{nix} is the probability of person n scoring X rather than X-1. This probability is modeled as a function of the ability β_n of person N and the difficulty δ_{ix} of the Xth step in item i. δ_{ix} is defined as the point on the ability scale where the probability of scoring x becomes greater than the probability of scoring X-1.

The two typical forms of equating with Rasch models are common item or common person equating. In the random groups design used in this study there are neither common items among test forms nor common

persons between groups. However, since the two groups are randomly equivalent any mean difference in logit ability between group A and B may be assumed to be due to mean difficulty differences between test forms Y and X. Therefore, if we wish to place the step values of form X on the Y scale we need to add to the X step values a linking constant, txy, given by

$$txy = \overline{X}_B - \overline{Y}_A \quad . \tag{4}$$

Method

This study used data from the Maryland Functional Writing Test (MFWT) to equate scores obtained from two prompts given in 1984 to the scores obtained from two different prompts administered in 1983. On the MFWT each student is required to write two passages. The first passage is in response to a prompt which is intended to elicit narrative writing. The second passage is in response to a different prompt which requires expository writing. Each passage is independently rated by two raters on a four-point holistic scale. The two ratiings on each passage are then averaged to provide a score on each writing domain (i.e., narration and explanation). The total test score is obtained by summing the two averages yielding a total score range of 2.0 to 8.0 in .5 intervals. All equating is done at the total test level where a passing score has been set at 5.5.

The data presented in this study were obtained during a field testing of many prompts (the 1983 and 1984 prompts were embedded in the field test design). The field test was conducted outside the state of Maryland (to help insure test security) on a population similar to that in Maryland.

In this study 351 examinees were administered two 1983 writing prompts, while 347 examinees were administered two 1984 writing prompts. Since these prompts were spiraled within classroom it may be assumed that the two groups of examinees were randomly equivalent in terms of ability.

Results

Linear Equating

The results of the linear equating are presented in Table 1. Table 1 also reports the raw scores, passing rates associated with each raw score, and various descriptive statistics for the 1983 and 1984 tests. In Table 1 the 1984 writing scores have been equated to the 1983 scores. Using equation 1 the slope of the linear equating line equals 1.07 with an intercept equal to -.59. The fifth column of Table 1 presents the equated raw scores. The sixth column of Table 1 shows the equated raw scores rounded off to the nearest .5 interval. This type of rounding must be done on the MFWT because the raw scores on the test can only be expressed in .5 intervals. The last column in Table 1 are the standard errors of equating based on equation 2. The results of the linear equating indicated that no change in score values were required within the score range from 5.0 to 8.0. However, within the score range from 2.0 to 4.5, an adjustment of .5 should be subtracted from the 1984 scores to make them comparable to the 1983 scores. Based on the linear equating and a passing standard of 5.5 there would be no change in the pass/fail rate in this particular set of data.

Estimates of the reliability (and therefore the standard errors of measurement) of the writing test may be obtained by procedures used by Coffman (1971, p. 282). The correlations between the two independent ratings for the 1983 narration and explanation prompts were .78 and .75, respectively. For the 1984 prompts they were .84 and .77. The average of the two 1983 correlations was .765, while the 1984 average was .805. Assuming that these average correlations represent the average reliability of the scores on single prompt, then an estimate of the reliability of the total test (based on the sum of the ratings on the two prompts) may be obtained by applying the Spearman-Brown formula. Using this procedure it was found that the reliability of the 1983 and 1984 scores were .87 and .89, respectively. Therefore, the standard errors of measurement for the two years were .56 and .46.

Polychotomous Rasch Model Equating

The results of the Rasch model equating are presented in Tables 2-5. Tables 2 and 3 show the step difficulties and standard errors obtained by fitting the partial credit model in equation 3 to the 1983 and 1984 ratings. On each passage the first and second ratings were treated as a separate item (yielding a four, rather than a two item test). This practice was an artifact of the scoring procedure used on the MFWT. The procedure was required to make the number of intervals in the logit ability metric equal the number of intervals in the total raw score scale. The adjusted step values in Table 3 were obtained by applying the linking constant given in equation 4. The mean logit ability obtained from the unadjusted 1983 and 1984 step difficulties were .84 and .94, respectively. According to equation 3 the 1984 step difficulties can be placed on the 1983 scale by adding .94 - .84 = .10 logit to each 1984 step difficulty.

Tables 4 and 5 show the raw score to logit ability transformation for the 1983 and 1984 prompts. The logit ability scores have also been linearly transformed into scale score units. The linear transformation uses an intercept of 318.48 and a slope of 20. This was done in order to make the passing raw score of 5.5 equal to 340 scale score units on the 1983 prompts (these were the prompts on which the standard was set). This is consistent with the other Maryland Functional multiple choice tests which have been calibrated with the dichotomous Rasch model and have a passing standard of 340 scale score units.

We see in Table 5, however, that a 5.5 on the 1984 test is associated with a scale score unit of 339 which is just below 340. We also see from Table 5 that the next highest raw score that exceeds 340 is a 6.0 with an associated scale score of 354. Based on the polychotomous Rasch model equating, in order to keep the standard of 340 constant from 1983 to 1984, the cut-score in the raw score metric would have to be raised from 5.5 to 6.0 on the 1984 prompts. We see from Table 1 this would result in 50.7 percent of students passing rather than 58.8 percent.

DISCUSSION

The results of the study suggest two conclusions. The first is that the effects of statistical equating depends in part on where within the score range the passing score happens to be. As we can see from the data in Tables 1-5, different passing scores would yield different results. For example, in this set of data, linear equating would not affect the passing score within the range of 5.0 to 8.0, but would alter the standard if it were within the 2.0 to 4.5 range.

The second conclusion is that different results are obtained with different equating methodologies. This should not be surprising since the same finding has occurred wth other recent psychometric undertakings (such as standard setting and item-bius procedures).

The major problem with statistically equating direct assessments of writing is related to the "discreteness" of the scale. With so few score points any adjustment will, in most cases, have uneven effects throughout the score range and may have drastic effects on pass/fail rates. The only solution to the problem is to score writing passages with a method other than holistic scoring. Whatever the method, it must yield a sufficient number of ratings to permit equating methodologies to operate. Although this may increase the cost of scoring and training time of judges, it will permit large scale testing programs to apply the same psychometric rigor to writing assessment as is currently applied to the testing of other basic skill areas.

In general, this author recommends that the equating of direct writing assessments be accomplished with the general polychotomous form of the Rasch model. There are several reasons for this recommendation. In the first place the method is probably as good or better than other methods. Secondly, the polychotomous model represents a natural extension for those testing programs already using the dichotomous Rasch model as the psychometric underpinning of multiple choice tests. Third, the software is available and easy to run. Fourth, the polychotomous model yields item parameters and test scores with the same invariance properties as the dichotomous model; (i.e., person scores are independent of writing prompts, and prompt difficulties are

independent of ability distributions). Finally, the polychotomous model will permit the banking of writing prompts and ultimately the choice of prompts based on test targeting strategies. In fact, all of the advantages of the Rasch model for multiple choice tests, potentially, may be realized in direct writing assessment.

Table 1

Raw Scores, and Percent Passing at Each Raw Score
For 1983 and 1984 Maryland Functional
Writing Tests, Level II

1983		1984				
Raw Score	Percent Passing	Raw Score	Percent Passing	*Raw Scores Based on Equating	Equated Raw Scores Rounded to Nearest .5	Standard Error of Equated Scores
8.0	5.7	8.0	10.1	7.97	8.0	.17
7.5	11.7	7.5	14.7	7.44	7.5	.15
7.0	21.7	7.0	29.1	6.90	7.0	.14
6.5	32.0	6.5	39.5	6.37	6.5	.12
6.0	46.5	6.0	50.7	5.83	6.0	.11
5.5	57.6	5.5	58.8	5.30	5.5	.11
5.0	68.7	5.0	72.9	4.76	5.0	.12
4.5	78.4	4.5	80.1	4.23	4.0	.13
4.0	91.5	4.0	91.6	3.69	3.5	.15
3.5	95.2	3.5	93.9	3.16	3.0	.17
3.0	98.0	3.0	96.8	2.62	2.5	.19
2.5	99.4	2.5	97.4	2.09	2.0	.22
2.0	99.7	2.0	99.7	1.55	2.0	.24

Descriptive Statistics

	1983	1984
Sample Size	351	347
Mean	5.5	5.7
Median	5.6	5.8
Standard Dev.	1.5	1.4
Skewness	-.147	-.399
Kurtosis	-.546	-.194
Reliability	.87	.89
Standard Error of Measurement	.54	.46

*Based on linear equating

Table 2

Partial Credit Step Difficulties and
Standard Errors for the Maryland
Functional Writing Test, Level II, 1983

Item Name	Step Values (Standard Errors)			Fit Statistic
	2	3	4	
Narration (Rater 1)	-2.89(.32)	.37(.17)	2.77(.18)	-1.25
Narration (Rater 2)	-3.05(.33)	.46(.16)	3.17(.19)	-1.33
Explanation (Rater 1)	-3.87(.41)	.68(.16)	2.98(.19)	-1.23
Explanation Rater 2)	-4.14(.44)	.75(.16)	2.77(.18)	-1.43

Table 3

Adjusted Partial Credit Step Difficulties
and Standard Errors for the
Maryland Functional Writing Test,
Level II, 1984

Item Name	Step Values (Standard Errors)			Fit Statistic
	2	3	4	
Narration (Rater 1)	-4.30(.63)	-0.14(.18)	2.44(.17)	- .44
Narration (Rater 2)	-3.84(.51)	-0.05(.18)	2.85(.18)	- .52
Explanation (Rater 1)	-2.36(.28)	1.00(.17)	3.67(.21)	-2.80
Explanation Rater 2)	-2.27(.28)	.68(.17)	3.51(.20)	-1.58

Table 4

Raw Scores, Logit Ability Transformation, Scaled Scores and Standard Errors For the Maryland Functional Writing Test, Level II, 1983

RAW SCORE	LOGIT ABILITY	LOGIT STD.ERROR	SCALED ABILITY	SCALED STD.ERROR
8.0	5.424	1.674	427	33
7.5	4.179	1.116	402	22
7.0	3.201	.897	382	18
6.5	2.453	.841	368	17
6.0	1.760	.827	354	17
5.5	1.076	.830	340	17
5.0	.366	.861	326	17
4.5	-.447	.949	310	19
4.0	-1.472	1.057	289	21
3.5	-2.566	1.015	267	20
3.0	-3.562	1.000	247	20
2.5	-4.689	1.173	225	23
2.0	-6.065	1.760	197	35

Table 5

Raw Scores, Logit Ability Transformation, Scaled Scores and Standard Errors For The Maryland Functional Writing Test, Level II, 1984

RAW SCORE	LOGIT ABILITY	LOGIT STD.ERROR	SCALED ABILITY	SCALED STD.ERROR
8.0	5.697	1.712	432	34
7.5	4.394	1.142	406	23
7.0	3.355	.936	386	19
6.5	2.537	.885	369	18
6.0	1.773	.868	354	17
5.5	1.028	.860	339	17
5.0	.285	.869	324	17
4.5	-.499	.905	308	18
4.0	-1.368	.956	291	19
3.5	-2.314	.986	272	20
3.0	-3.326	1.033	252	21
2.5	-4.543	1.214	228	24
2.0	-6.017	1.821	198	36

References

Angoff, W.H. Scales, norms, and equivalent scores. In Robert L. Thorndike (ed.). Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Coffman, W. E. Essay Examinations. In Robert L. Thorndike (ed). Educational Measurement. Washington, D. C.: American Council on Education, 1971.

Masters, G.N. A Rasch model for partial credit scoring. Psychometrika, 1982, 47, 149-174.

Masters, G.N. Constructing an item bank using partial credit scoring. Journal of Educational Measurement, 1984, 21, 19-32.

Phillips, Gary W., Mead, Ronald J., and Ryan, Joseph P. Use of the general polychotomous form of the Rasch model as a means of calibrating and equating indirect and direct assessments of writing achievement. A paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, 1983.

Williams, P.L. Practical applications of item response theory in large-scale testing programs: equating direct writing assessments. A paper presented at the 12th annual Large-Scale Assessment Conference, Boulder, Colorado, 1983.

Wright, B.D. and Masters, G.N. Rating Scale Analysis. Chicago: MESA. Press, 1982.